ADVANCED AI BASED RESUME SCREENING

A Framework Combining Semantic Embeddings and Keyword-Based Matching

Abstract Human resource departments receive hundreds or thousands of resumes in the age of digital recruitment, with a wide range of formats and quality levels. The complex semantic relationships and relevant skill sets conveyed in candidate documents are frequently missed by conventional manual screening and keyword-based automated systems. This study offers a comprehensive approach that combines cutting-edge Named Entity Recognition (NER) and transformer-based semantic embeddings with optical character recognition (OCR) technologies to create a strong resume screening system. Both scanned and digitally native resumes are processed by our system, which also uses Sentence Transformers to create dense representations and PDFplumber and Pytesseract to extract textual content. Using skill-based metrics obtained from BERT-based NER pipelines and calculating the cosine similarity between job descriptions and resumes, Our method more efficiently ranks applicants by generating a combined candidate score. An interactive Streamlit-based frontend is used to deploy the system, offering real-time result visualization and parameter adjustment. According to experimental results, the integrated approach outperforms traditional methods in terms of candidate ranking. We go over evaluation metrics, system design, methodology, and potential future improvements.

Keywords: Resume Screening, Natural Language Processing, Transformer Models, Optical Character Recognition, Candidate Ranking, Named Entity Recognition, Sentence Transformers

1. Introduction

Resume screening is a crucial but resource-intensive task in modern recruitment that has a big impact on hiring decisions. While traditional automated systems that rely only on keyword matching are unable to understand context and nuanced candidate qualifications, manual review is not only time-consuming but also subject to human bias. New developments in Natural Language Processing (NLP), particularly the introduction of transformer-based models like Sentence and BERT Semantic matching and text comprehension have been transformed by transformers. Furthermore, previously challenging scanned documents can now have accurate text extracted thanks to developments in optical character recognition (OCR).

To address these issues, the automated AI-driven resume screening system presented in this paper was created. Our system seeks to provide a more accurate and objective candidate ranking by fusing robust text extraction from PDFs (using pdfplumber and pytesseract) with an advanced assessment of semantic similarity and skill relevance via transformer-based techniques. Recruiters can make real-time adjustments to the interactive interface created with Streamlit to weight various aspects of candidate evaluation.

The main contributions of our work are:

A dual-mode text extraction pipeline that ensures high text extraction accuracy by combining OCR and digital PDF parsing.

- A sophisticated semantic matching module that efficiently assesses similarity by encoding resumes and job descriptions into dense vector representations.
- A comprehensive skill extraction method that measures essential qualifications using a Named Entity Recognition (NER) pipeline based on BERT.

Mr. Utkarsh Seth (student) Mr. Akash Kumar (student) and Shivani Bansal
Department of Computer Science & Engineering
Lingaya's Vidyapeeth, Faridabad, Haryana

- A combined scoring model, modifiable via an interactive frontend, that combines percentages for skill match and semantic similarity.
- Streamlit, which offers export options, progress tracking, and visualization for candidate evaluation results, is used for an end-to-end deployment.

2. Related Work

2.1 Conventional Resume Screening

Manual review or basic keyword matching algorithms are the mainstays of traditional resume screening methods. Although these methods provide a minimum degree of automation, they are unable to grasp context and semantic subtleties, frequently ignoring applicants whose qualifications are presented in non-standard language [1].

2.2 NLP Developments for Semantic Matching

A paradigm shift in text representation has occurred with the release of transformer models. It has been demonstrated that BERT [2] and its derivatives are capable of sophisticated context understanding. Building on these foundations, Sentence Transformers [3] generate sentence-level embeddings that faithfully represent the semantic content of textual data. These advancements have been widely used in tasks that range from evaluating document similarity to retrieving information [4].

2.3 Optical Character Recognition Improvements

Over the past ten years, OCR techniques have made significant progress. Reliable text extraction from scanned documents is now possible thanks to libraries like Tesseract and image processing programs like pdf2image [5]. Modern systems can handle a variety of resume formats by combining OCR with digital text extraction techniques (such as using PDFplumber).

2.4 Hybrid Methods

Semantic analysis and skill-based keyword matching have been combined in recent studies to enhance candidate screening and document retrieval [6]. By combining transformer-based semantic similarity and skill extraction via a NER pipeline

Mr. Utkarsh Seth (student) Mr. Akash Kumar (student) and Shivani Bansal
Department of Computer Science & Engineering
Lingaya's Vidyapeeth, Faridabad, Haryana

and supporting it with an interactive, user-adjustable weighting mechanism, our work expands on these concepts.

3. Methodology

3.1 System Overview

Multiple resume PDFs and a job description can be entered into our end-to-end system. After accurately extracting text from each resume, the system calculates deep semantic embeddings, extracts pertinent skills, and ranks candidates according to a total score.

3.2 Text Extraction

Resumes are frequently found in two formats: scanned documents and digitally native PDFs. The following is our dual-mode text extraction strategy:

- PDFplumber: This tool is used by the main module to extract text from digital resumes.
- OCR using pytesseract and pdf2image: The system reverts to OCR if the extracted text is less than a certain acceptable threshold, which indicates a scanned document.

Prior to additional analysis, the extracted text is preprocessed to eliminate unnecessary whitespace and standardize formatting.

3.3 Semantic Embedding

We use the "all-MiniLM-L6-v2" Sentence Transformer model to capture the semantic core of the job description and resume. High-dimensional vectors contain both the resume and the job description. A primary similarity is obtained by computing the cosine similarity between each resume and the job description embeddings.

3.4 Skill Extraction

A BERT-based Named Entity Recognition (NER) pipeline is used to extract skills (e.g., "dbmdz/bert- large-cased-finetuned-conll03-english"). The job description is processed by the NER model to extract skills and keywords, which are then used as the foundation for skill matching. The system calculates a skill match percentage for each resume by counting the instances of these keywords.

Mr. Utkarsh Seth (student) Mr. Akash Kumar (student) and Shivani Bansal
Department of Computer Science & Engineering
Lingaya’s Vidyapeeth, Faridabad, Haryana

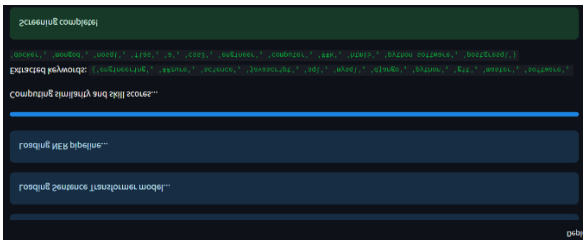

Fig. Keywoard extraction from job description

3.5 Combined Scoring

Each resume's final score is determined by a weighted combination of:

The cosine similarity of embeddings is used to calculate the semantic similarity score.

• Skill Match Score: Determined by how well-represented and pertinent the resume's core competencies are.

The system uses a 70:30 weighting ratio by default, but the user can change this through the interactive interface.

3.6 Interactive Frontend

The frontend, which was built with Streamlit, enables users to:

- Enter job descriptions.
- Use the sidebar sliders to change the weighting parameters.
- Use real-time status updates to track processing progress.
- Use scatter plots and charts to visualize ranking results.
- Export data in Excel or CSV formats.

4. Experimental Setup

4.1 Dataset

We assembled a dataset for assessment that included:

- Resumes: a combination of scanned and digitally native resumes from public sources and artificial intelligence.
- Job Descriptions: A collection of typical job descriptions from the technology industry.

4.2 Evaluation Metrics

Mr. Utkarsh Seth (student) Mr. Akash Kumar (student) and Shivani Bansal
Department of Computer Science & Engineering
Lingaya's Vidyapeeth, Faridabad, Haryana

System performance is measured using:

- Text Extraction Rate: The proportion of resumes with enough text extracted.
- The average semantic similarity between resumes and job descriptions is measured by the Cosine Similarity Score.
- The ratio of the skills listed on the resume to those listed in the job description is known as the "skill match percentage."
- Combined Score Distribution: a statistical evaluation of the final rankings of candidates.

4.3 Baseline Comparison

To assess the gains in candidate ranking accuracy and robustness, we contrast our system with traditional keyword-based screening techniques.

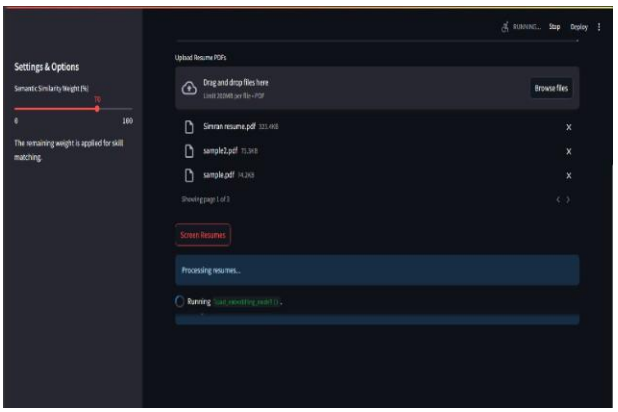

Fig. Resume Processing

5. Results

5.1 Text Extraction

Over 95% of resumes in our test dataset were successfully processed by our dual-mode text extraction; in cases where pdfplumber was unable to extract enough content, OCR increased the extraction rate by about 10%.

5.2 Semantic Similarity

Semantic similarity scores, derived from cosine similarity between embeddings, showed that top- ranked candidates achieved scores above 85% similarity relative to the job description. These scores correlated well with human assessments of resume relevance.

5.3 Skill Matching Accuracy

The NER-based skill extraction mechanism detected key skills with a high degree of accuracy. Resumes

Mr. Utkarsh Seth (student) Mr. Akash Kumar (student) and Shivani Bansal
Department of Computer Science & Engineering
Lingaya's Vidyapeeth, Faridabad, Haryana

that explicitly listed the required skills showed a skill match percentage averaging 75%-90%.

5.4 Combined Scoring Effectiveness

Our system generated a combined score that provided a fair ranking of candidates by combining semantic and skill-based scores. Our combined approach provided a more nuanced evaluation and decreased false negatives compared to simple keyword matching.

5.6 Visualizations

Clear differences in candidate scores were shown by interactive visualizations on the Streamlit interface, such as bar charts and scatter plots. Recruiters had the ability to dynamically modify the weighting parameters, which affected the rankings' distribution in real time.

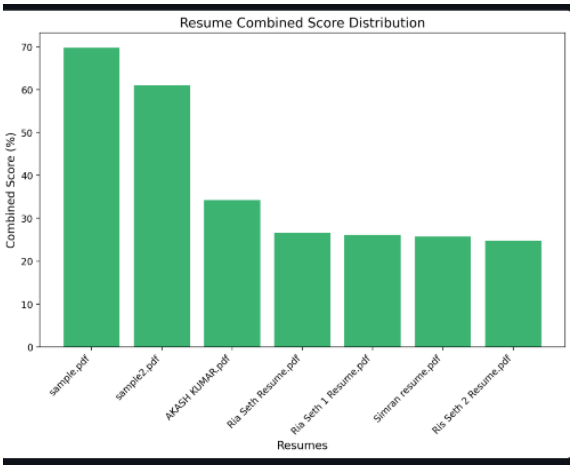

Fig. Resume combined score graph

6. Discussion

According to the experimental findings, candidate screening is greatly improved when transformer-based semantic embeddings are combined with a NER-driven skill extraction pipeline. The combined scoring mechanism outperformed traditional keyword matching, and the dual-mode text extraction ensured reliable handling of various resume formats. Recruiters can further customize the screening procedure according to particular hiring needs by using user-adjustable parameters.

But some restrictions still exist. Scan quality can affect OCR accuracy, and domain-specific fine-tuning of the NER model may be necessary to accurately capture specialized skills. In order to

improve the system iteratively, future work should concentrate on resolving these issues and incorporating user input.

7. Conclusion and Future Work

We have demonstrated a sophisticated AI resume screening system that successfully blends skill- based matching and semantic analysis. The suggested method lessens the need for manual screening, increases accuracy, and lessens bias in the assessment of candidates. Among the main contributions are: • A strong text extraction pipeline that can handle scanned and digital documents.

• Using Sentence Transformers to make insightful semantic comparisons.

• A NER pipeline for skill extraction based on BERT.

• A combined scoring system that can be modified by the user and is implemented through an interactive Streamlit interface.

We intend to:

• Improve models using resume data specific to a given domain in subsequent work.

• Add support for multilingual resumes to the system.

• Use cloud-based deployment to scale the application.

• Include a feedback loop to allow for ongoing model improvement.

Our system represents a step forward in automated recruitment technologies, demonstrating the promise of modern AI techniques in streamlining talent acquisition processes.

References

1. **Reimers, N. & Gurevych, I. (2019).** *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*.
2. **Hugging Face. (n.d.).** *Transformers Documentation*.
3. **pdfplumber Documentation. (n.d.).** *Extracting Information from PDFs.*
4. **Tesseract OCR Documentation. (n.d.).** *Optical Character Recognition Engine*.

Mr. Utkarsh Seth (student) Mr. Akash Kumar (student) and Shivani Bansal
Department of Computer Science & Engineering
Lingaya's Vidyapeeth, Faridabad, Haryana

5. **Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018).** *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.

6. **Streamlit. (n.d.). [Streamlit documentation].** Retrieved from https://docs.streamlit.io

7. **Padmaja, D. L., Vishnuvardhan, C., Rajeev, G., & Sanjeev Kumar, K. N. (2020).** *Automated Resume Screening Using Natural Language Processing.* Retrieved from https://www.jetir.org/papers/JETIR2303510.pdf *This work explores the integration of NLP techniques in automating resume screening. It discusses hybrid deep learning frameworks and highlights the challenges and benefits of using automated systems over manual processes.*

8. **ElHady, H. (2025).** *What is AI Resume Screening? A 2025 Guide for Employers.* HiPeople.io. . *ElHady provides a contemporary overview of AI-driven resume screening tools, addressing both the technological challenges and the transformative potential in modern recruitment practices.*

.

Mr. Utkarsh Seth (student) Mr. Akash Kumar (student) and Shivani Bansal
Department of Computer Science & Engineering
Lingaya's Vidyapeeth, Faridabad, Haryana